FERMILAB-CONF-10-343-CD

Distributed Monte Carlo production for DZero

Joel Snow for the DØ collaboration

Department of Mathematics, Langston University, Langston, OK, USA

E-mail: snow@fnal.gov

Abstract. The DZero collaboration uses a variety of resources on four continents to pursue a strategy of flexibility and automation in the generation of simulation data. This strategy provides a resilient and opportunistic system which ensures an adequate and timely supply of simulation data to support DZero's physics analyses. A mixture of facilities, dedicated and opportunistic, specialized and generic, large and small, grid job enabled and not, are used to provide a production system that has adapted to newly developing technologies. This strategy has increased the event production rate by a factor of seven and the data production rate by a factor of ten in the last three years despite diminishing manpower. Common to all production facilities is the SAM (Sequential Access to Metadata) data-grid. Job submission to the grid uses SAMGrid middleware which may forward jobs to the OSG, the WLCG, or native SAMGrid sites. The distributed computing and data handling system used by DZero will be described and the results of MC production since the deployment of grid technologies will be presented.

1. Introduction

Simulation or Monte Carlo (MC) data is crucial to physics analysis. At Fermi National Accelerator Laboratory (FNAL) [1] the Tevatron luminosity and hence raw data volume are at record levels. This presents challenges for both analysts and production. While at the same time, personnel and computing resources have been migrating to the experiments at the Large Hadron Collider (LHC) [2]. How does a mature experiment continue to produce quality physics results in an ever-changing environment which requires it to produce more with less? DZero [3] has employed a successful two pronged strategy based on increased automation and the leveraging of resources and support.

2. DZero Experiment

DZero took Run 1 data from 1992 to 1996. The detector underwent a substantial upgrade from 1996 to 2001, and has been running nearly continuously since then. The upgraded Run II 4500 metric ton DZero detector measures 10m by 10m by 15m. It consists of approximately one million data channels and is able to inspect 1.7 million proton - antiproton collisions per second. Data is recorded at about 100 events per second with a data flow of 20 Megabytes per second. In a year 300,000 Gigabytes of data are recorded. Through November 2008 the DZero experiment has recorded 4.5 billion events.

The DZero experiment is a global enterprise consisting of 550 scientists, 150 graduate students from 19 countries on four continents and 90 institutions, 38 of these in the U.S. The experiment will run through 2010 with an expected data set increase of 50-100%. This represents a particularly challenging environment as resources migrate to LHC experiments.

DZero is a mature experiment but it remains nimble. DZero has a history of adopting innovative technologies, in particular the distributed data handling system “Sequential data Access to Metadata” (SAM) [4, 5, 6, 7, 8, 9], and early adoption of the grid for production using SAMGrid [10, 11]. The experiment has a significant investment in these technologies. DZero has pursued access to additional resources by aggressively using grid technology which allows opportunistic resource usage. DZero is able to mix “traditional” dedicated and opportunistic resources. Additional resources are made available by utilizing grid interoperability. For DZero this leverages resources and support, reducing personnel needs per CPU hour.

2.1. SAM

In 1997, prior to grid deployment, FNAL and DZero created the SAM distributed data handling system first used in production by DZero. It consists of a set of servers working together to store and retrieve files and metadata. SAM has permanent storage and local disk caches. Its database tracks file replica locations, file metadata, and job processing history, as well as other information used by the experiments. SAM delivers files to jobs using GridFTP over the Wide Area Network (WAN) and also provides job submission capabilities which are not used in MC production.

2.2. SAMGrid

SAMGrid is a FNAL developed grid that predates OSG deployment. It was first used by DZero for global MC production in 2004. SAMGrid is the synthesis of SAM and “Job and Information Management” (JIM) [12] components. It provides the user with transparent remote job submission, data processing, and status monitoring. SAMGrid is Virtual Data Toolkit (VDT) [13] based, using the Globus Toolkit [14] and Condor-G [15] components.

Logically SAMGrid consists of multiple execution sites, a resource selector, multiple job submission (scheduler) sites, and multiple clients (user interfaces) to the submission site. For SAMGrid operation the user submits a job request to the queuing node, based on the Condor scheduler, via a remote client based on Condor client commands. Jobs are matched and submitted to execution sites which are based on the Globus gatekeeper/jobmanager. At the execution site job requests are split into multiple job instances. For MC jobs there are 250 events per job. Job instances are submitted to a local batch queue or to another grid. The execution site controls data traffic shaping and triggers data delivery, including executable binaries, control scripts, environment data, and input file data. See Figure 1.

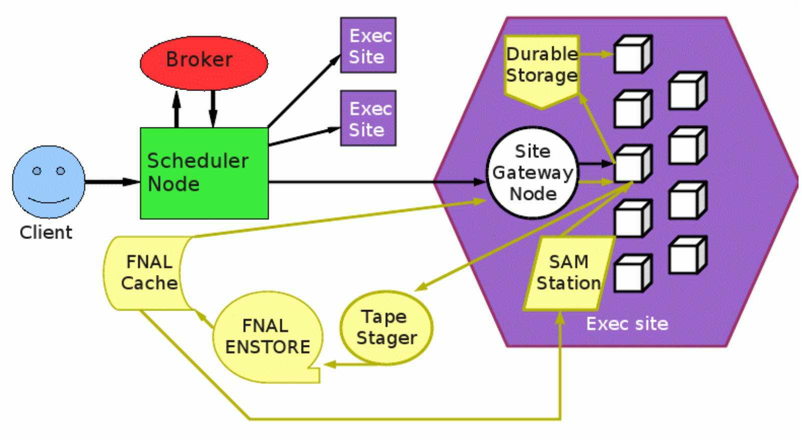

Figure 1. SAMGrid components.

2.3. SAMGrid Interoperability

As the Open Science Grid (OSG) [16] and Worldwide LHC Computing Grid (WLCG) [17] became operational it was desirable to leverage these resources for DZero. FNAL and DZero developed and deployed SAMgrid interoperability [18] with both WLCG and OSG resources. In this scenario the execution site acts as a Forwarding node to the other grid. The Forwarding node packages SAMGrid jobs for OSG/WLCG submission via Condor-G/Glite. See Figure 2.

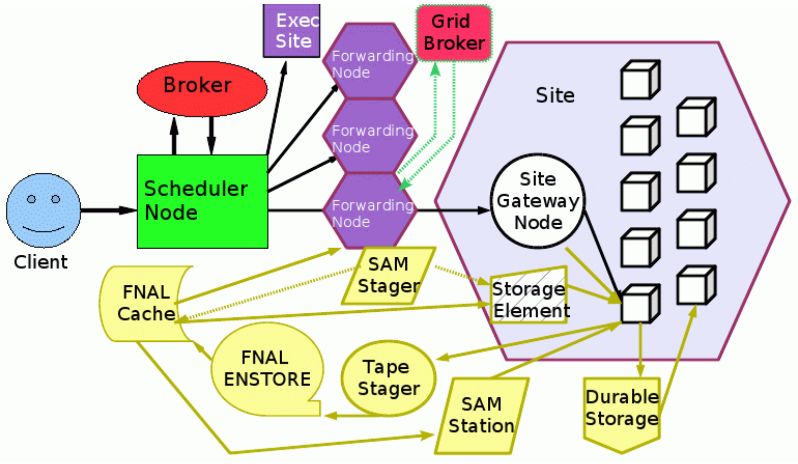

Figure 2. SAMGrid OSG/WLCG components.

2.4. Automation, Consolidation, Exploitation

SAMGrid deployment provided an immediate and substantial increase in production by enabling the use of remote opportunistic and dedicated resources. Experience showed that operation of SAMGrid for production required substantial operations personnel and expert support. At the same time personnel and FNAL and remote site support have been migrating to the LHC experiments. To deal with this situation DZero decided to increase automation, reduce the number of SAMGrid sites, and increase use of the OSG and WLCG. This gains the experiment people power by automating tasks and reducing needs for operations. The increase in use of other grids has the double benefit of coming with their own support structure and providing opportunistic job slots.

3. Production System

The DZero MC production system gets work from the SAM request system. Physics groups' MC requests are parameterized and prioritized. A request is in the form of a Python dictionary stored in SAM.

3.1. MC Applications

A typical request has four phases running one application each. The first phase is the generator phase which simulates the physics of interest. This phase produces a file which becomes the input data for the second phase. The simulation phase propagates the particles of interest through the detector. This phase also produces a file which becomes input for the next phase. The digitization phase puts the simulated data in the form of raw data from the detector and overlays it with a generic background. The output file of that phase becomes the input to the final reconstruction phase, which reconstructs the event with the first pass analysis code. The metadata of all phases are saved in SAM, but typically only the output file of the last phase is eventually stored on tape at FNAL.

3.2. MC Grid Job Flow

The job's bootstrap executable (3 MB) is delivered from the execution site/forwarding node (ES/FN) via Condor-G. The bootstrap executable unpacks a gridftp client which fetches 20 MB of initial environment and utility files from the ES/FN. At this point a SAM client environment wrapping the gridftp transport is available. The job uses the SAM client to download 800 MB of applications and the execution environment from SAM cache located locally or remotely to the job in the worker's sandbox. An optional input data file (200 Mb - 1 GB) may also be downloaded at this time is required by the job. These jobs start at the simulation phase and do not run the generator phase. Zerobias overlay files (∼300 MB) are also downloaded at this time to the job from SAM cache. Events from the overlay files are randomly merged with simulated physics events during the digitization phase. For OSG or WLCG jobs *no VO specific pre-installed software is required at the job site.* The output data file produced by the reconstruction phase (15-30 MB) is stored in a "durable location" known to SAM for later handling.

SAMGrid jobs are broken into 250 events chunks at the ES/FN for submission to the batch/Condor-G/Glite system. This is an execution time trade-off to maximize usable sites. Output file size for 250 events is too small for efficient tape storage so these files are merged together in a separate grid job. The 10,000 event merged files (1 GB) are stored on tape via SAM and the unmerged files are deleted.

3.3. Automatic MC Request Processing

The automatic MC request processing system (AutoMC) handles requests from initial approved status in SAM to final data storage in SAM. It is easy to use and minimizes manpower needs. AutoMC is site independent. It can submit jobs destined for any of the three grids used by DZero (SAMGrid, OSG, WLCG). AutoMC handles recovery of common failure modes and integrates with the pre-existing MC request priority protocol.

The AutoMC system has five components, (1) a SAM client for database queries; (2) JIM for job submission and monitoring; (3) a daemon which periodically awakens to do work; (4) a local database to store request processing data and history; and (5) grid credentials. The daemon has three components. The first is a "Broker" which assigns requests to resources. It submits the initial job of the request, is configured on the fly via a text file, and can pass site decision making to other brokers *e.g.* the OSG Resource Selection Service (ReSS) [19]. The second daemon component is the "Running Request Monitor" which finishes the production and merging of a request by submitting appropriate recovery jobs. The third component is the "Request Status Publisher" which keeps track of processing details in log files, and creates a web page of an updated status summary of AutoMC.

3.4. Data Transport Issues

Native SAMGrid jobs use LAN transport to workers from the reliable SAM station caches. Jobs forwarded to OSG and WLCG use WAN transport to workers from remote SAM caches. This is an advantage since no VO specific software needs to be pre-installed at job site and provides great site selection flexibility. However this is also a disadvantage since WAN transport has been observed to be less reliable than LAN transport. Less than optimum efficiencies result. For OSG sites the use of Storage Elements (SE's) as SAM caches mitigates the problem. SE's are presently in use at Indiana University, Michigan State University, Purdue, University of California at San Diego, and University of Nebraska at Lincoln. A significant improvement in efficiency is observed at these sites.

3.5. Production System Resources

The MC production system uses a variety of dedicated and opportunistic resources, both large and small, on four continents. These include a single large non-grid site at CCIN2P3 [20] in Lyon, France. This site is very productive, efficient, and provides great flexibility in MC job processing. There are five legacy native SAMGrid sites operating, one in the Czech Republic, two in Germany, one in the US, and one in China. DZero has used WLCG resources including Computing Elements (CE's) in France, the Netherlands, and the United Kingdom, and SAMGrid-WLCG infrastructure in those countries plus Germany. DZero has used OSG resources including CE's and SE's, and SAMGrid-OSG infrastructure in the United States and Brazil. Resources range in size from small clusters with less than 10 cores, to large sites where fair share is twenty batch slots or less, to large clusters where up to 2000 MC jobs may be running on a good day. The variety of resources has proven a strength as production continues when a segment has a problem.

4. MC Production Results

Figure 3 shows weekly production results and the cumulative total of events since September 2005 (when latest MC epoch began) for the calendar year 2008. The production is measured as merged events stored in SAM. The period averaged 15.5 million events per week and totaled 805 million events for the year.

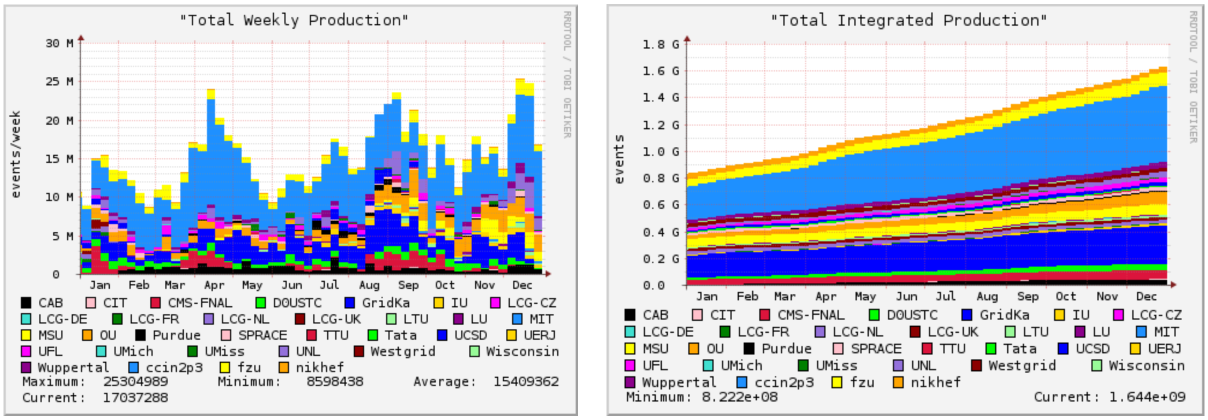

Figure 3. (a) Weekly production results and (b) cumulative total of events since September 2005 for calendar year 2008.

Tables 1 & 2 show the characterization of production by segment and geographic region respectively for 2008. WLCG production was down for most of 2008 due to hardware problems and available personnel issues. About three-fifths of production is done using grid technology and two-thirds of production is done in Europe.

Table 1. Characterization of production by segment for 2008 in millions of events.

Segment	Events (M)	Average/week	%
Non-grid	324.4	6.24	40.3
OSG	262.8	5.04	32.6
SAMGrid	212.9	4.02	26.4
WLCG	5.0	0.11	0.6

Table 2. Characterization of production by geographic region for 2008 in millions of events.

Region	Events (M)	%
Europe	483	64
N. America	232	31
Asia	34	4
S. America	11	1

Despite the lack of WLCG production the number of produced events since September 2005 almost doubled during 2008. To fully appreciate the success of DZero's three pronged strategy of (a) automation coupled with (b) flexible, diverse, opportunistic and dedicated resource utilization, and (c) the leveraging of the support services of the grids used, one needs to look on a longer time scale than one year. Figure 4 shows weekly production results and the cumulative total of events since September 2005 for the three years spanning 2006-2008. The three year period averaged 10.0 million events per week and totaled 1.5 billion events. The dip in production during the first five months of 2007 was due to the diversion of MC resources to do data reprocessing [21]. Clearly evident from the cumulative graph is the acceleration of the MC production rate. Several factors contributed to this acceleration including the increasing power of machines and networks. More importantly the rate increase was the result of more resources made available through opportunistic grid usage. In 2006 SAMGrid deployment was becoming more widespread after the production proof of concept in 2004 and initial deployment in 2005. In 2006 OSG forwarding was used for the first time in a production environment and WLCG forwarding was first deployed for production. In 2007 OSG usage became widespread and WLCG production grew more than making up for the diversion of resources to reprocessing. In 2008 WLCG production faltered but OSG production more than made up for it as more resources became available and efficiency improved. Non-grid production also had a significant increase in 2008 as more newer and faster cores were added to the existing facility.

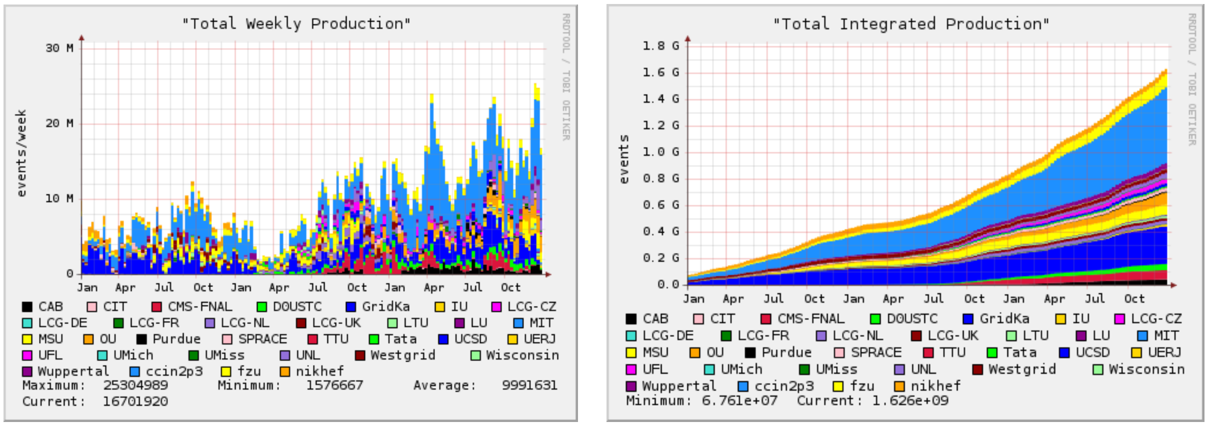

Figure 4. (a) Weekly production results and (b) cumulative total of events since September 2005 for the period 2006-2008.

The impact grid usage has had on DZero MC production can be seen in Figure 5 and Tables 3 & 4 which show MC event and data production by segment for the previous five years. In 2004 the first deployment of SAMGrid occurred and gives an idea of pre-grid production levels.

SAMGrid quickly grew to be a significant fraction of overall production surpassing non-grid production in 2006. Forwarding to other grids became a significant source of production in 2007. In 2008 SAMGrid production was deprecated while both OSG and non-grid production increased dramatically.

Table 3. MC event production by segment in millions of events for the previous five years.

Period	Total	Non-grid	SAMGrid	OSG	WLCG
2007/12/26-2008/12/26	794.8	315.6	213.6	259.7	5.8
2006/12/26-2007/12/26	398.2	109.1	158.1	96.5	34.4
2005/12/26-2006/12/26	348.0	144.4	195.5	0.5	7.6
2004/12/26-2005/12/26	98.1	68.6	29.5	0.0	0.0
2003/12/26-2004/12/26	42.4	41.8	0.6	0.0	0.0

Table 4. MC data production by segment in terabytes for the previous five years.

Period	Total	Non-grid	SAMGrid	OSG	WLCG
2007/12/26-2008/12/26	67.8	26.9	18.4	22.0	0.5
2006/12/26-2007/12/26	31.6	7.3	13.2	8.2	2.9
2005/12/26-2006/12/26	23.0	9.4	13.1	0.0	0.5
2004/12/26-2005/12/26	6.0	4.1	1.9	0.0	0.0
2003/12/26-2004/12/26	1.9	1.9	0.0	0.0	0.0

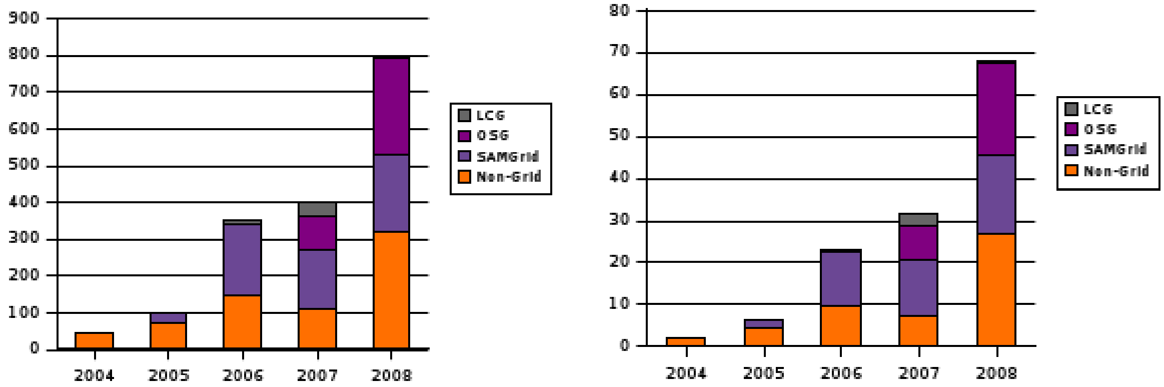

Figure 5. MC production in (a) millions of events and (b) terabytes of data by segment for the previous five years.

5. Future Prospects

In the future non-grid and SAMGrid resources will not expand appreciably if at all. As the LHC experiments come on line, diminishing job slots on opportunistic resources are expected. What can be done to maintain production rates? Not all possible OSG resources have been utilized. Restoration of SAMGrid forwarding to WLCG by using the OSG-WLCG gateway is underway, which may provide an additional 25% of production capacity. Are other grids usable? Work has started on exploring job forwarding to the TeraGrid [22] via an OSG-TeraGrid gateway at FNAL. It is also appropriate to ask how well existing resources are being used. Changing data transport patterns through the use of local SE's at OSG sites led to a significant increase in efficiency. Similar improvements in efficiency by infrastructure expansion and tuning may be possible.

6. Conclusion

DZero's early adoption of innovative technologies has dramatically increased MC production. DZero was the first deployment of the SAM distributed data handling system which first gave MC production a global reach. DZero was the first to deploy SAMGrid for production in 2004. DZero has leveraged the use of OSG and WLCG resources through interoperability with SAMGrid since 2006. DZero was the first opportunistic usage of OSG SE's for production in 2008. DZero is now exploring the use of the TeraGrid through interoperability. DZero has deployed a automated MC production system since 2007.

In conclusion, DZero MC production is doing well as the production rate doubled in the last year despite no appreciable WLCG contribution. Although DZero probably faces future resource diminution, efforts are underway to help sustain production. These include increasing opportunistic computing and storage, increasing job efficiency, minimizing faulty requests reaching the system, increasing the infrastructure's resiliency to failure, and exploring new resources via grid interoperability. It is expected MC production will not be a limiting factor in any DZero physics analysis for the remaining life of the experiment.

Acknowledgments

We thank all the members of the DØ collaboration who have contributed to the Monte Carlo effort through the years. We thank the Fermilab Computing Division for the products, tools, and support that are critical to the success of DØ 's Monte Carlo production effort. We thank the members of the OSG and WLCG consortia for access to their resources and support.

References

[1] The Fermi National Accelerator Laboratory home page, `http://www.fnal.gov`

[2] The Large Hadron Collider home page: `http://lhc.web.cern.ch/lhc/`

[3] The DZero collaboration The DØ upgrade: the detector and its physics *Fermilab Pub-96/357E*
The DZero experiment home page: `http://www-d0.fnal.gov`

[4] The SAM project home page: `http://d0db.fnal.gov/sam/`

[5] White V*et al* DØ data handling *Proc. of Computing in High-Energy and Nuclear Physics (CHEP01)* Beijing China Sep. 2001 (New York: Science Press) pp 20-8

[6] Carpenter L *et al* SAM overview and operational experience at the DZero experiment *Proc. of Computing in High-Energy and Nuclear Physics (CHEP01)* Beijing China Sep. 2001 (New York: Science Press) pp 310-4

[7] Carpenter L *et al* SAM and the particle physics data grid *Proc. of Computing in High-Energy and Nuclear Physics (CHEP01)* Beijing China Sep. 2001 (New York: Science Press) pp 724-9

[8] Carpenter L *et al* Resource management in SAM - the DØ particle physics data grid *Proc. of Computing in High-Energy and Nuclear Physics (CHEP01)* Beijing China Sep. 2001 (New York: Science Press) pp 730-5

[9] Lueking L *et al* The data access layer for DØ run II *Proc. of Computing in High-Energy and Nuclear Physics 2000* Padova Italy Jan. 2000 (Padova: Imprimenda) pp 462-6

[10] Terekhov I *et al* 2002 Meta-Computing at DØ *Nuclear Instruments and Methods in Physics Research* Section A, NIMA14225, **502/2-3** pp 402-6

[11] Garzoglio G *et al* 2002 The SAM-GRID project: architecture and plan *Nuclear Instruments and Methods in Physics Research* Section A, NIMA14225, **502/2-3** pp 423-25

[12] Terekhov I *et al* Grid job and information management for the FNAL Run II Experiments *Proc. of Computing in High-Energy and Nuclear Physics (CHEP2003)* La Jola Ca. USA Mar. 2003

[13] The Virtual Data Toolkit home page: `http://vdt.cs.wisc.edu/`

[14] The Globus Toolkit home page: `http://www.globus.org/`

[15] The Condor Project home page: `http://www.cs.wisc.edu/condor/`

[16] The Open Science Grid home page: `http://www.opensciencegrid.org/`

[17] Apostolakis J *et al* Architecture blueprint requirements technical assessment group (RTAG) *Report of the LHC Computing Grid Project* CERN OCT. 2002
The Worldwide LHC Computing Grid home page: `http://lcg.web.cern.ch/`

[18] Garzoglio G, Baranovski A and Mhashilkar P The SAM-Grid/LCG interoperability system: a bridge between two grids *Proc. of Computing in High-Energy and Nuclear Physics (CHEP2006)* Mumbai, India Feb. 2006 (New Dehli: Macmillan India) pp 677-80

[19] The OSG Resource Selection Service web page:
`https://twiki.grid.iu.edu/bin/view/ResourceSelection/WebHome`

[20] The CCIN2P3 web page: `http://cc.in2p3.fr/cc_accueil.php3?lang=en`

[21] Abbott B, Baranovski A, Diesburg M, Garzoglio G, Kurca T and Mhashilkar P 2008 *J. Phys.: Conf. Series* DZero data-intensive computing on the Open Science Grid **119**
`http://www.iop.org/EJ/article/1742-6596/119/6/062001/jpconf8_119_062001.pdf`

[22] The TeraGrid home page: `http://www.teragrid.org/`